# A Visual and Statistical Analysis of Various Covid-19 Data

Austin Way[1, a]

*Undergraduate Statistics, Student ID: 914741141*

(Dated: 9 June 2021)

This document analyzes the effects of COVID-19 on the United States using a variety of different statistical methods. Visualization is performed on various aspects of demographic and vaccination data to provide suggestions to the public. Mathematical methods are also performed on the data to outline some of the underlying mechanisms that drive a viral outbreak. While these methods may not be used in the field of epidemiology, they are effective, data-driven methods that have the ability to uncover meaning from complex, multi-dimensional data in various different fields of study.

[a] arway@ucdavis.edu

## I. Introduction

In late December of 2019, news began circulating of a ”pneumonia of unknown origin” in Wuhan, China. By the end of January 2020, the World Health Organization had declared this new disease a ”global health emergency”, and by early March it would be officially classified as a pandemic. The novel coronavirus disease, or COVID-19 (CO for corona, VI for virus, D for disease) is caused by a virus called SARS-CoV-2. This virus belongs to a family of viruses that causes diseases ranging from the common cold to nastier ones such as SARS (Severe Acute Respiratory Syndrome) and MERS (Middle East Respiratory Syndrome).

Data science plays an important role in battling COVID-19. From visualizing data in order to inform the public, to fitting models in order to analyze trends and predict outbreaks, the importance of accurate and publicly available data cannot be overstated. Throughout the pandemic, the U.S. Centers For Disease Control and Prevention (CDC) has been posting weekly updates and providing many publicly available data sets in order to keep the public informed about the COVID-19 situation in the United States. As a data scientist, I will be looking at some of this data in order to find interesting relationships and answer questions about the mechanisms driving the outbreak in the United States.

Some questions of interest include: How are the hardest-hit states' vaccination efforts going? Are certain age groups / sexes / other population groups affected more by the pandemic than others? What steps can the country take to continue to reduce cases of COVID?

## II. Data

For my research, all data directly associated with COVID-19 was obtained from the U.S. Center for Disease Control's publicly available data[1]. Outside data was retrieved from various sources, cited in Appendix A. Below is a description of each of the COVID datasets used for analysis:

### COVID Data Sets

- **Case Surveillance Data**[2]

  This dataset contains 12 variables measured over $25,607,582$ observations. Each observation represents a separate case. The variables of interest are the date the case was reported, the date symptoms onset (if applicable), the status of the case (laboratory confirmed or probable), the sex and age group of the patient, whether or not they went to the hospital, whether or not they died, and whether or not the patient had pre-existing conditions. This is by far the largest dataset used in this analysis, and is used primarily to demonstrate patterns in case distributions of different groups.

- **Cases and Deaths by State**[3]

  This dataset contains 15 variables measured over $30,062$ observations. Each observation represents a single day for a single jurisdiction. The variables of interest include the date, state, total cases in that state, new cases from the previous day, total deaths in that state, and new deaths from the previous date. This type of data is useful for measuring the impact of the pandemic in each of the U.S. states/territories.

- **Moderna Vaccine Allocations by Jurisdiction**[4]

  This dataset contains 4 variables measured over $1,575$ observations. Each observations represents a given week for a single jurisdiction. I will focus on all four variables: week of allocations, jurisdiction, dose 1 allocations, and dose 2 allocations. This data will be combined with the Pfizer data to measure the number of doses allocated to each of the jurisdictions. It is useful for measuring vaccine hesitancy when partnered with the vaccine distribution data.

- **Pfizer Vaccine Allocations by Jurisdiction**[5]

  This dataset's structure is identical to the Moderna dataset above, however, Pfizer vaccines began being allocated 1 week earlier than the Moderna vaccines, resulting in this data having $1,638$ observations. This data will be combined with the Moderna data to measure the number of doses allocated to each of the jurisdictions.

- **Vaccinations In the U.S. by Jurisdiction**[6]

  This dataset contains 69 variables and $11,584$ observations. Each observation represents a single day for a single jurisdiction. The variables of interest include distributed Moderna and Pfizer vaccines, the total amount distributed per 100k population, the number of administered Moderna and Pfizer vaccines, as well as data on the number of first doses and second doses, and series completion for the vaccines (both doses administered). This data is useful for comparison with the allocation data, as well as for analyzing the effect of vaccination on each state's COVID cases.

### Supplementary Data Sets

Two supplementary data sets were required to assist in the analysis of COVID data. The State Abbreviations[7] dataset allowed the conversion of abbreviations to state names to assist with visualization, and the State Populations[8] dataset allowed the addition of state populations to the case data.

### Data Cleaning and Combination of Data Sets

To prepare the data for analysis, different methods of preprocessing were applied to each of the data sets. Below are listed various preprocessing steps. Note that some preprocessing was applied to combat the memory limitations of my personal computer, and should not be repeated given an adequately powerful computer.

- **Cleaning of Case Surveillance Data:** From preliminary analysis of the values for each column, it was noted that the *Deaths* and the *Preexisting Conditions* columns both had over 10 million rows with ”Unknown” or NA values. Removing the empty values from both of these columns results in an overall reduction from 25,000,000 to 2,183,094 rows. Although this removal was to reduce computational intensity, it will help with subsequent analysis as well. To further assist in the analysis, the removal of missing values was applied to the *Hospital*, *Race/Ethnic Group*, and *Sex* variables. After all removal of missing values, the dataset contained 1,499,404 rows.
- **Preprocessing of By-State COVID Data:** To help with visualization, the first step was to convert the *code* column to show full state names instead - this was

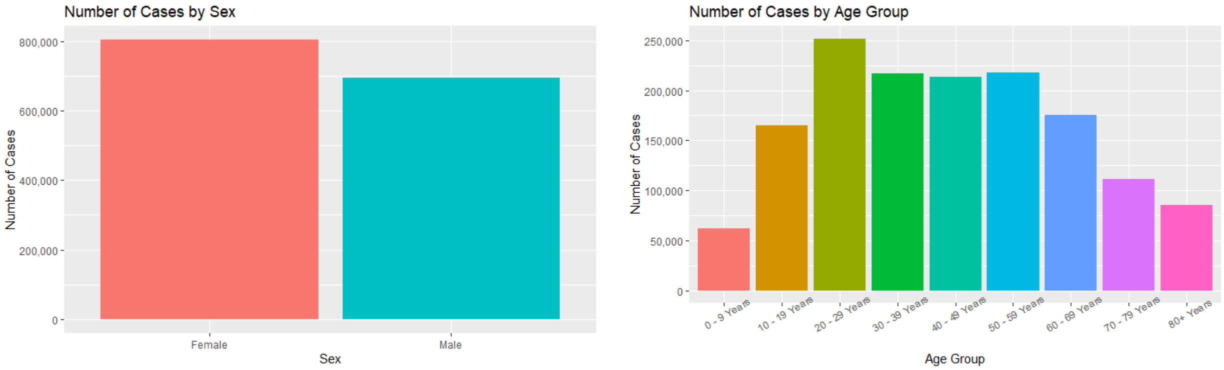
performed by combining the dataset with the state abbreviation dataset. Next, the *Date* column was converted from a class POSIXCT(a time object) to be of class Date, for easier combination with the vaccine data. An extra column ”WeekOf” was added as well, for easy combination with weekly data (as opposed to daily).

– **Preprocessing and Combination of Vaccine Data:** The first, and easiest, step of processing the vaccine data was combining the Pfizer and Moderna Allocation data, as they are structured exactly the same. A column was then added, ”TotalAlloc”, to define the total number of doses allocated for each observation. A second column ”WeekInEffect” was then added by adding 5 weeks to Pfizer allocations and 6 Weeks to Moderna allocations, to define the week the allocated dose would go into effect (to be used later in analysis). The combined Pfizer and Moderna allocation data was then added as column ”TotalAllocPerWeek” in the larger distribution and administration dataset.[?]

# III. Data Visualization

## Case Data by Group

To begin analyzing COVID-19 data, it would be helpful to get an understanding of the underlying distributions of COVID cases. Below are plots of cases by Sex and Age Group:



(a) (b)

FIG. 1. Cases by Sex(a) & Age Group(b)

The plot of cases by sex shows that there does not seem to be any significant relationship between the number of cases and sex of the patient, as it matches the standard U.S. population distribution[9]. The age group plot almost matches the standard distribution of age as well, except that the values of 0-9 years and 10-19 years are much lower than in the U.S. population. Compared to 31% of the population being under 18, COVID-19 cases of children 0 - 19 only make up 15% of all cases. Studies have shown that COVID-19 does not affect young children nearly as badly as adults[10], however they are still susceptible to the virus. The lack of cases in the data is likely due to most juvenile COVID-19 cases not presenting with symptoms, and therefore going undetected. Also, adults age 60+ only represent 16% of the population while making up 24% of the total number of cases.

Next, we look at the breakdown of US deaths by case:

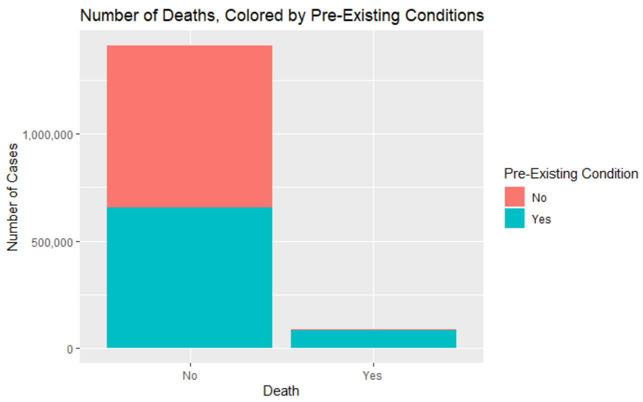


FIG. 2. Number of Deaths (Yes or No)

The figure shows that the overwhelming proportion of cases did **not** result in death, however, in the cases that did, almost all of them had pre-existing conditions. This is a very significant proportion, and people with these conditions should take extra measures to protect themselves.

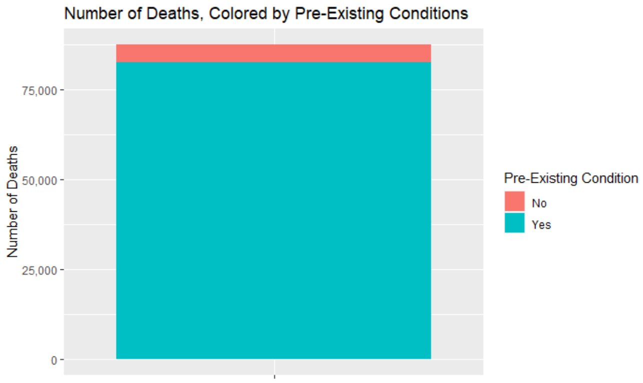


FIG. 3. Breakdown of Deaths

Of the observations of the dataset, a whopping 94.25% of deaths were in patients with pre-existing conditions.

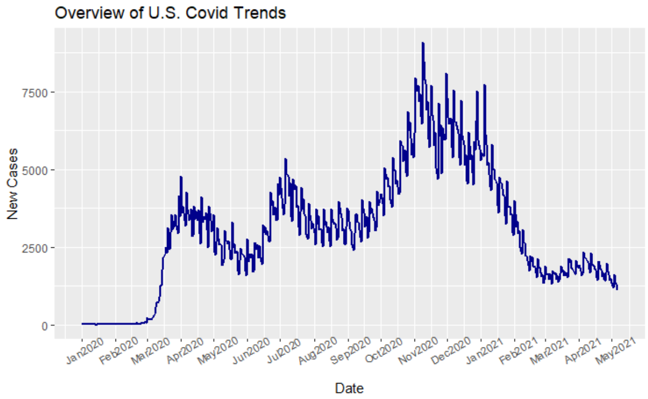


FIG. 4. Overview of U.S. COVID infections

### Case Data by Time

As shown in Figure 2, there were several small peaks during the outbreak of COVID in the United States. The initial surge was in early March 2020, hitting an early peak. As businesses began gradually reopening and people began to go out more (against recommendations), another peak was observed in early July 2020. The largest peak was observed in late 2020/early 2021, as a few states went against the CDC guidelines and opened up completely.

### Vaccines & Other Preventative Measures

In order to fight the spread of the virus and protect the people, many preventative measures were put into place over the course of the pandemic. Some of these measures included mandating face masks, shutting down non-essential businesses, and, eventually, vaccine allocation and distribution. On March 19, 2020, California became the first state to issue a statewide ”Stay At Home Order” as well as a statewide mask mandate, leading the charge for other states to do the same[11]. The results of these measures are reflected in the overall case trend, creating the dip after the first peak. As well as stay-at-home orders and mask mandates, vaccination has been crucial in combating COVID-19. As vaccinations began rolling out, the U.S. saw the ever-growing number of cases begin to decline nationwide.

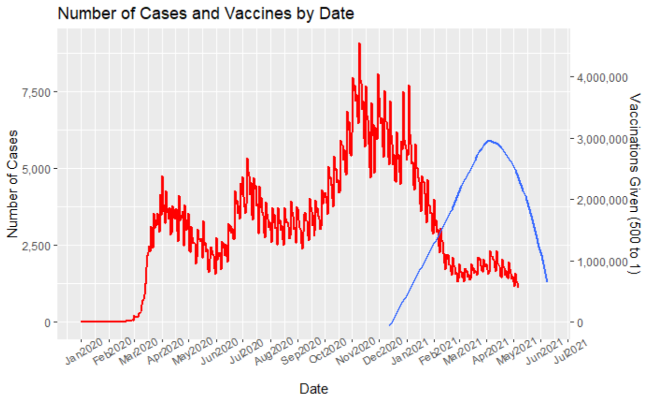


FIG. 5. COVID cases (red) and Vaccinations (blue)

Note: The # of vaccinations has been reduced by a factor of 500 for visualization

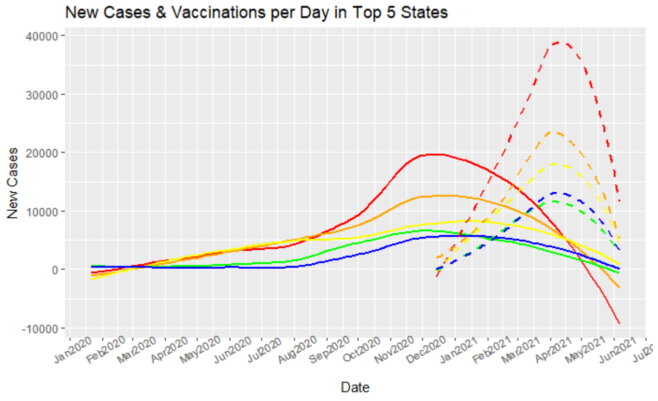


FIG. 6. The five highest-case states' vaccination efforts

Some states, however, were unable to efficiently distribute vaccines immediately - whether it be due to infrastructural limitations or to vaccine hesitancy of the population. In Table I, we examine the to-date levels of cases, deaths, and vaccinations by population in the five states with the highest cases per population.

| State | Cases/Pop | Deaths/Pop | Mortality | Vaccinations/Pop |
|---|---|---|---|---|
| North Dakota | 0.146 | 0.0020 | 0.014 | 0.328 |
| Rhode Island | 0.143 | 0.0026 | 0.018 | 0.504 |
| South Dakota | 0.142 | 0.0023 | 0.016 | 0.374 |
| Utah | 0.129 | 0.0007 | 0.006 | 0.351 |
| Tennessee | 0.127 | 0.0018 | 0.014 | 0.306 |

TABLE I. Most Cases by Population and Vaccination Efforts

Although these states have had the most relative cases, they do not seem to be far off from the U.S. average of 37.8% vaccination rate. Potentially, the vaccine distribution in these states was delayed, but with this information, we cannot attribute the inflation rate of infections to lack of vaccination.

## IV. Mathematical Analysis

The science of epidemiology is very complex, and without proper knowledge of the mechanisms driving infections, it is hard to understand outbreaks. In this section, a linear model of cases by vaccinations and Principal Component Analysis are applied to some of the COVID data to help create some understanding of the relationships between different aspects of an outbreak.

### Linear Model

From Figures 5 and 6 above, we can see that cases begin to drop significantly at the onset of vaccination administration. To test if vaccination has a statistically significant impact, we fit a simple linear model:

$$Y = \beta_0 + \beta_1 * X_1 \tag{1}$$

Where:

$$\beta_0 = \text{Expected \# of New Cases (daily)}$$

$$\beta_1 = \text{Coefficient of Vaccinations}$$

$$X_1 = \text{Number of New Vaccinations (daily)}$$

Since we are modeling the effect of vaccinations, we limit the data to January 2021 - present, as there were no vaccinations administered prior. Fitting the model to the data, we get the equation:

$$NewDailyCases = 895 - 0.0132 * DailyVaccinations \tag{2}$$

This model returns a p-value very close to 0 (<.00001), and as such, is statistically significant at 99% confidence. There is therefore a proven link between the number of daily administered vaccinations and the number of daily cases. The $R^2$ value of this linear model is .325, indicating that vaccines are responsible for explaining 32.5% of the change in new cases. This is a large amount of the variance to be explained by a single variable, and further exemplifies the effect of vaccinations.

From the results above, it is easy to see that vaccinations have decreased the number of cases in the US. The other measures that have been implemented, i.e. mask mandates, stay-at-home orders, etc., have effects that are not so easy to prove. In the next section, mathematical analysis will be performed on various time points of the data.

## Principal Component Analysis (PCA)

Principal Component Analysis is a statistical technique used widely in multivariate data analysis. At its core, PCA is a dimensionality reduction, reducing the data into *principal components*. Each of these components is a linear combination of the variables in the data (with formulas referred to as *loadings*), and are responsible for explaining a certain proportion of the variance in the data. Inserting the values of each row into the formula will return a *score* for each component.The loadings and scores of these components can then be analyzed to find interesting relationships in the data.

To analyze COVID data, PCA was performed on two separate groups: the earliest cases in the U.S. (March 15th, 2020) and the most recent available data (June 6th, 2021), and on the highest peak of the pandemic (January 7th, 2021) and the height of the first peak (April 01, 2020). The variables used in this PCA are New Cases, Total Cases, New Deaths, Total Deaths, # of Administered Vaccines, Mortality Rate (# Deaths / # Cases) and Cases by Population (# Cases / State Population). Each row of data measures the data for a single U.S. state on the day of analysis. For March 15th and April 1st data, the Administered Vaccines variable is excluded, as there were no vaccines yet produced and therefore no useful information would be provided. The desired number of components will be the smallest number with cumulative variance of at least 90%.

### *1. March 15, 2020 & June 6, 2021*

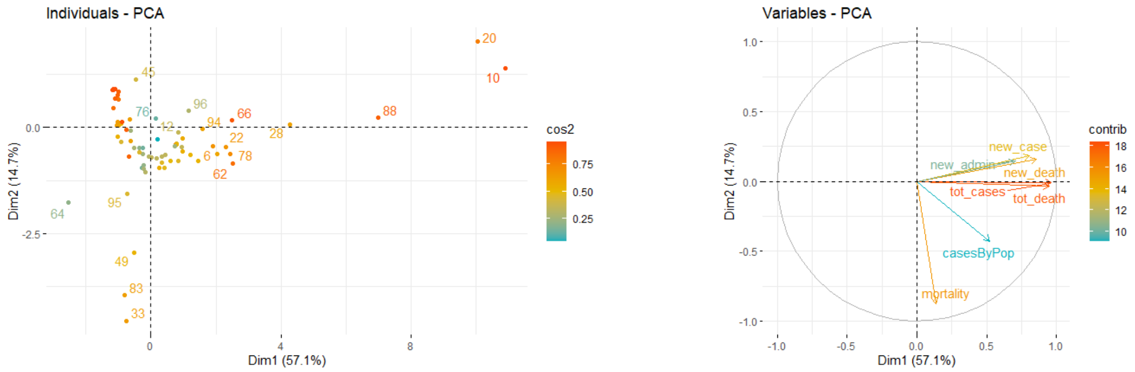
| | Comp. 1 | Comp. 2 | Comp. 3 | Comp. 4 |
|---|---|---|---|---|
| St.Dev | 1.999 | 1.014 | 0.937 | 0.797 |
| Cum. Variance | 0.571 | 0.718 | 0.844 | 0.934 |
| Loadings: | Comp. 1 | Comp. 2 | Comp. 3 | Comp. 4 |
| New Case | 0.404 | | 0.333 | 0.375 |
| Total Cases | 0.481 | | | |
| New Deaths | 0.429 | | 0.303 | 0.321 |
| Total Deaths | 0.474 | | | |
| Administered Vacc. | 0.352 | | | -0.830 |
| Mortality | | -0.862 | 0.490 | |
| Cases by Population | 0.262 | -0.422 | -0.727 | 0.223 |

TABLE II. PCA Table

For most Americans, March 15 was the very beginning of the pandemic, as shutdowns and mask mandates began to go into effect around the country. Now, in early June 2021, it feels like the end of the pandemic is nearing as those mandates are being lifted. In March 2020, the number of cases was low and the number of deaths even lower. In June 2021, the number of new cases are the lowest in months, after battling through the worst of COVID-19.

In Table I note the captured proportion of variance for each component - for this PCA, just four components were needed to explain 90%. Component 1 (PC1) is a combination of all variables except for Mortality. Note the positive correlation between these six variables, as all coefficients are positive. The states with the highest scores for this component are likely to be the states worst affected by the pandemic, as cases, deaths, and vaccinations are high. For example, in Figure 5, point 10 and point 20 correspond, respectively, to the data for California and Florida on June 6, 2021. The states with the lowest scores for this component are the states that had yet to have had an outbreak in March 2020, as they will have the lowest values for cases, deaths, and obviously, vaccinations. There is a notable amount of clustering around the origin, as many observations of states at each of these time points are seeing similar numbers of new cases and new deaths. The idea behind this component is fairly intuitive, as it is clear that a change in time from the start to the end of the pandemic will result in an increase of total cases and vaccinations.

Component 2 details a positive correlation between mortality and cases by population. This indicates that the more cases a state has, the higher the mortality rate. Remembering that the data is a combination of the start and end of the outbreak, it follows that the observations from the latter group will have both higher mortality rates and total cases. Therefore, this component does not give meaningful insight into the data, but rather explains what is assumed to be true. Component 3 is interesting, indicating that the number of new



(a)

(b)

FIG. 7. Plot of Individual Scores and Variable Correlation

cases, new deaths, and the overall mortality have a negative correlation to the number of cases by population. As the number of overall cases by population increase, it should be expected that there is a decrease in mortality (as states learn to battle the virus) and a decrease in the number of new deaths and cases per day. Lastly, Component 4 details a negative relationship between administered vaccines against new cases and deaths, as well as overall cases by population. The states with the highest scores for this component will have successfully implemented a vaccination program, and are beginning to stop the outbreak for good.

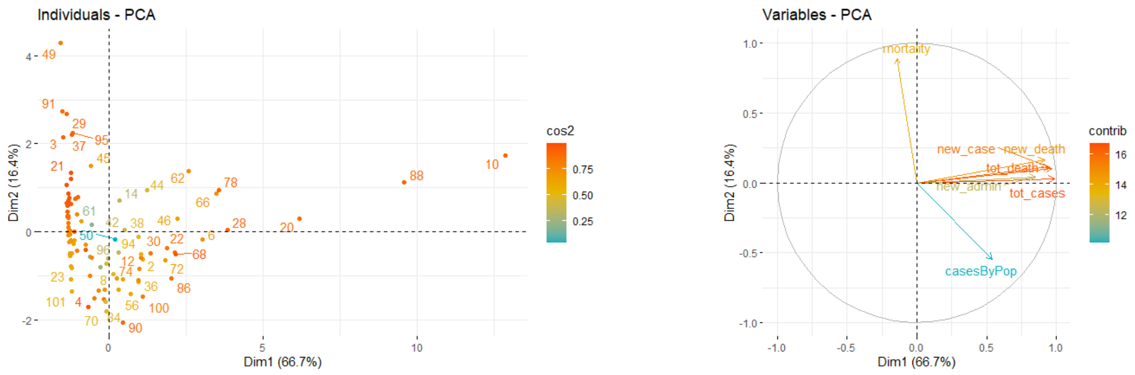
Principal Component Analysis of the March 2020 and June 2021 data has demonstrated some of the driving forces behind changes in data throughout the outbreak in the country. Almost all of the variance in the data can be explained by four components, indicating the severity of the outbreak at each observation, the mortality and total case difference from the beginning to the end of the pandemic, indicating the success of each state in fighting the virus, and indicating the success of each state in their vaccination programs. Now that the low ends of the COVID-19 curve have been analyzed (Fig. 2), let us perform PCA on two of the few single-day peaks.

### *2. April 1, 2020 & January 7, 2021*

| | Comp. 1 | Comp. 2 | Comp. 3 |
|---|---|---|---|
| St.Dev | 2.161 | 1.071 | 0.795 |
| Cum. Variance | 0.667 | 0.831 | 0.921 |
| Loadings: | Comp. 1 | Comp. 2 | Comp. 3 |
| New Case | 0.439 | | 0.202 |
| Total Cases | 0.456 | | |
| New Deaths | 0.423 | | |
| Total Deaths | 0.447 | | |
| Administered Vacc. | 0.391 | | |
| Mortality | | 0.832 | -0.541 |
| Cases by Population | 0.250 | -0.510 | -0.797 |

TABLE III. PCA Table

April 01, 2020, was the first peak of cases in the United States, and some would argue that fear of the virus was highest at this point. After months, however, whether due to ignoring guidelines or just the natural course of the virus, the highest single-day case peak was on January 7, 2021 - a whopping 843,000 new cases[12]. At this time, vaccinations had just begun being allocated and would soon be distributed to each of the states. For the measurements from these two peaks, the data needed to be reduced to just three components, explaining 92.1% of overall variance. Figure 8 below visualizes the scores of each observation for the first two components, and the correlation of the variables for each component.



(a)

(b)

FIG. 8. Plot of Individual Scores and Variable Correlation

From Table II, Component 1 is very similar to the first component in the previous section. It is a positive combination of all variables (excluding mortality) and represents the natural growth of the outbreak. Component 2 is opposite of the second component in the previous section, representing a strong negative correlation between mortality and cases by population. These measurements were taken at the high points of the pandemic, and therefore the mortality increase from the surge of new cases would not go into effect (given time for course of virus in those individuals). The final component describes a relationship such that as the number of new cases increase, the overall mortality rate and cases by population decrease.

From component 2, the relationship between the change in cases at the peaks (high numbers of new cases) and a decrease in mortality (total deaths / total cases) is understood, however it is important to note the negative correlation between new cases and cases by population seems to be over-fitting the variance in the data, as there is no intuitive negative relationship between the two variables. A potential real-world explanation could be that at these time points, the number of new cases was very high in states with low numbers of total cases by population, resulting in this interesting relationship.

The factors that influence the spread of a virus are complex and often require the usage of complex, multi-stage models. PCA does not tell the entire story, but rather gives some insight to help facilitate understanding of an outbreak. From the starting and ending data (section IV 1), high levels of clustering were seen, indicating that most states were in similar situations at those times. Section IV 2 shows a large amount of clustering as well, with more significant differences among points on both axes (Fig. 6(a)) Some states at these peak points were dealing with a much larger influx of cases and deaths, and therefore had higher values for Component 1. Overall, it has been shown that the principal changes in data at peaks and at lows have some commonalities, like having almost identical Component 1's. They also have differences - at the low points, mortality and total cases by population are positively correlated, while at the high points they have negative correlation.

## V. Results & Findings

In Section III, the distribution of cases by Sex and Age were visualized. There was no significant relationship found between the number of cases and the sex of the patient, however, it was shown that patients age 0 to 19 only represented  15% of cases, when they make up  31% of the population. Adults over the age of 60 represented  24% of cases, although only making up  16% of the population. From these proportions, it is safe to assume that children are less susceptible to the virus, while older adults are more susceptible. It is important to note that this result does not imply children having immunity; the data is comprised of *reported* cases and it is likely that children and healthy young adults may not show symptoms and therefore not report their cases at the same frequency. Lastly, over 94% of deaths in the United States have been in people with pre-existing conditions as defined by the CDC.

In Section IV, a linear model was fitted to the data (exclusively from 2021) to demonstrate the effect of vaccinations on overall cases. The model returned a p-value of $>0.0001$, and was therefore significant. The $R^2$ value of this regression was .325, indicating that vaccines are responsible for explaining 32.5% of the change in new cases.

Later in the section, Principal Component Analysis was applied to two separate groups: a group of two low points (the start and end of the case distribution) and a group of high points (the first peak and the absolute peak). Using these analyses, the similarities and differences between the two groups were highlighted. It was noted that both time points share variance between states in the severity of their outbreaks, and differences in the relationships between total cases and mortality rate.

## VI. Conclusion

COVID-19 has left an unprecedented impact on the United States and the world. Thanks to data collected and updated by the Centers for Disease Control, data scientists have been able to process and visualize data to inform the public and find interesting relationships. In this report, various techniques were used to present and analyze data from multiple different sources. These techniques gave some insight into the mechanisms driving the outbreak, and how to reduce cases for good.

As the end of the pandemic seemingly draws nearer, the states must continue to keep battling the virus with vaccinations and other safety measures until it is deemed safe. Older adults (60+) should continue to be safe, even after vaccination, as they are the most susceptible populace group. Another group that should continue to be vigilant include those with pre-existing conditions, as an overwhelming proportion of deaths were in individuals with these conditions.

Overall, it is exciting to see the rate of infections dropping further and further. Hopefully, soon, COVID-19 will no longer be a threat to people around the world, and life can go back to normal.

## VII. Acknowledgements

In the absence of a group, I would like to thank my friends for taking time out of their finals schedules to help proofread my final draft:

Prabhjyot Mann, Aiken Tong, and Avinash Vadlamudi

## APPENDIX A: DATA SOURCES & REFERENCES

[1] “Data — Centers for Disease Control and Prevention” , https://data.cdc.gov/browse (Accessed on 06/07/2021).

[2] “COVID-19 Case Surveillance Public Use Data Profile — Centers for Disease Control and Prevention” , https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-Profile/xigx-wn5e (Accessed on 06/07/2021).

[3] “United States COVID-19 Cases and Deaths by State over Time — Centers for Disease Control and Prevention” , https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36 (Accessed on 06/07/2021).

[4] “COVID-19 Vaccine Distribution Allocations by Jurisdiction - Moderna — Centers for Disease Control and Prevention” , https://data.cdc.gov/Vaccinations/COVID-19-Vaccine-Distribution-Allocations-by-Juris/b7pe-5nws (Accessed on 06/07/2021).

[5] "COVID-19 Vaccine Distribution Allocations by Jurisdiction - Pfizer — Centers for Disease Control and Prevention" , https://data.cdc.gov/Vaccinations/COVID-19-Vaccine-Distribution-Allocations-by-Juris/saz5-9hgg (Accessed on 06/07/2021).

[6] "COVID-19 Vaccinations in the United States,Jurisdiction — Centers for Disease Control and Prevention" , https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc (Accessed on 06/07/2021).

[7] "List of state abbreviations" , https://worldpopulationreview.com/states/state-abbreviations (Accessed on 06/07/2021).

[8] "US State populations - 2018 — Kaggle" , https://www.kaggle.com/lucasvictor/us-state-populations-2018 (Accessed on 06/09/2021).

[9] "Age and sex composition: 2010" , https://www.census.gov/prod/cen2010/briefs/c2010br-03.pdf (Accessed on 06/09/2021).

[10] "Covid-19 (coronavirus) in babies and children - mayo clinic" , https://www.mayoclinic.org/diseases-conditions/coronavirus/in-depth/coronavirus-in-babies-and-children/art-20484405 (Accessed on 06/09/2021).

[11] "A timeline of covid-19 developments in 2020" , https://www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020 (Accessed on 06/09/2021).

[12] "Covid live update: 175,070,388 cases and 3,774,227 deaths from the coronavirus - worldometer" , https://www.worldometers.info/coronavirus/?fbclid=

IwAR1c8GJsTIjla-8FtG2OOmj8FHuYXhbw_6nhbeFxbnLQINMdyof0kDnGlXc (Accessed on 06/09/2021).